Is it Interesting?: Comparing human and machine judgements on the PETS dataset

Hogg, David C.; Dee, Hannah

Publication date:
2004

Citation for published version (APA):
Hogg, D. C., & Dee, H. (2004). *Is it Interesting?: Comparing human and machine judgements on the PETS dataset*. 49-55. http://hdl.handle.net/2160/5871

Is it interesting?
Comparing human and machine judgements on the PETS dataset

Hannah Dee & David Hogg
School of Computing
University of Leeds
Leeds LS2 9JT, United Kingdom
{hannah}{dch}@comp.leeds.ac.uk

Abstract

This paper presents a novel approach to evaluating the detection of unusual or interesting events in videos involving certain types of human behaviour, such as pedestrian scenes. The holy grail of computer vision for surveillance can be thought of as an interesting or unusual event detector which when given an input video stream, outputs some form of alarm whenever anything unusual happens inside its field of view. This paper addresses the question of how we would go about evaluating such a system, suggests one possible evaluative schema, and presents an example of this evaluative procedure in use on a prototype Interesting Event Detector.

1 Introduction

When you monitor a pedestrian scene, a number of different behaviour patterns can be observed. People walk along pathways and cars are driven along roads, and occasionally people will take shortcuts or get into a car or stop for a chat. Very occasionally, someone will do something different or *interesting* – something that does not fit our general understanding of what behaviour people exhibit in that scene. Humans are very good at detecting such events, but are not so good at articulating what exactly it is that makes such events unusual or interesting. It is worth making the distinction between events which are *interesting* and events which are *atypical* - most Computer Vision systems for surveillance attempt to detect the latter, whereas humans are much more interested (by definition) in the former. It is easy to imagine a system which would ring an alarm if a pedestrian strayed from a path. But if the path was blocked, this behaviour, although atypical, would not really be interesting.

A number of systems have been constructed which can be portrayed as attempts at identifying such events. Many systems are actually designed to do something else, and atypicality detection emerges as a "bonus" feature: by modelling a particular feature of the environment (path usage, patterns of pedestrian motion over time etc.) and determining which instances *do not* fit the model, some form of interesting-event detector has magically been constructed.

One approach is exemplified in [4], in which the typicality or otherwise of pedestrian trajectories is assessed based upon learned models of absolute location and speed over time. In [6], a model of the paths within a scene is constructed based upon the behaviour of pedestrians, and this path model can subsequently be used to detect unusual trajectories. The relationships between objects can also be used to judge typicality [7]. In [9] patterns of activity are learned at a site, and unusual event detection is performed by spotting events which do not fit the pattern or co-occurrence data. In [3] "suspicious" behaviour is correlated with the rapid head movements.

Determining the overall effectiveness of algorithms of this type has historically been unsystematic. This is acknowledged by the authors of [9], who state they are working on methods of evaluating the unusual event detection aspect of their work.

Evaluative techniques at their simplest involve investigating the problematic cases by hand - looking at the outliers - and saying "Yes, that's unusual" [9, 4]. One model, trained on pedestrians, had a major outlier which turned out to be a cyclist. This is, of course, confirmation that the model provides a reasonable basis for the detection of strange pedestrian activity, however, the confirmation such evidence provides is at best anecdotal. It is also completely self-justifying - if we look at the examples which do not fit the model, and find they are odd in some way, then of course they are interesting to us - they fit our frame of reference (or rather, they don't fit our frame of reference) by definition.

Another means of evaluating such systems is through the

use of "actors" [1]. These people are recorded behaving in an unusual fashion, and the system in question is evaluated on its ability to single out the sequences featuring the strangely behaving actors [3, 7, 4]. Problems with this approach are manifold, but all hinge upon the question of *whose idea of interesting or unusual we are dealing with*. If the decision as to what constitutes unusual behaviour is left up to the actors, questions about who the actors are, what their preconceptions of the project are and most importantly, their links to the software designers, become paramount. If the actors are lab-mates of the paper author, do they know how the algorithm in question works? The alternative case, where the actors are instructed by the system designer on the nature of unusual behaviour, could be even worse - it is easy to imagine a scenario in which the instruction "We need some footage of supicious behaviour, like walking from car to car across the car park in a wavy line" is issued. This is not exactly *good science*.

Computer Vision systems for surveillance are generally model based. And things which do not fit the model can only be classed as unusual or interesting with respect to that model. We cannot really claim that events which fall outside the model are interesting or unusual - all we can really say about them is just that they don't fit the model. Thus we really cannot claim any more or less for these interesting event detectors until we have a more principled way of evaluating their performance. This paper proposes a way out of this model-based trap - by providing a form of "ground-truth" for interestingness.

2 Am I interesting or not?

Within the surveillance domain, what we are interested in are events which might be associated with criminal or dangerous behaviour. A recent study [10] investigates whether such events can be predicted from CCTV footage - that is , whether it is possible to distinguish sequences where a crime was about to occur from neutral sequences. The authors conclude that not only is it possible, but that naïve observers perform as well as trained security guards. This suggests that there is no learned or innate ability to detect the type of events security guards detect.

Our central assumption is that benchmarking against a number of humans is an improvement over relying on the author, actors, or serendipity to provide some measure of the interestingness or otherwise of the data set.

The evaluative schema we propose involves requiring a number of volunteers (in this case, undergraduate and postgraduate students with no knowledge of the project being evaluated) to rank the behaviour of each agent in the scene in question. To assist in this task, separate videos are produced for each agent containing only those frames of video encompassing the agent's trajectory. A highlight indicates exactly the agent we are interested in - this makes the cognitive task of those evaluating much easier in scenes with multiple, occluded agents.

Volunteers are asked to rate the "interestingness" of these videos on a scale of 1 to 5. The instructions given to the volunteers were as follows:

> *"If you were a security guard, would you regard the behaviour of the agent highlighted in this video as interesting? Please indicate on the following questionnaire, with one being uninteresting and five being interesting."*

Volunteers were also invited to note down any comments they wished to make about any of the videos.

An average of the scores from the human rankers is then assumed to provide a simple measure of "interestingness": we choose the median, as this is less sensitive to outliers. We can then compare it directly to the output of any machine generated indication of typicality, and if we want our system to output a binary decision (interesting, or not) we can use ROC graphs to assist in the determination of a threshold.

However, the median is just one statistic we can use: the advantage of having the opinions of a number of people is that there is a richness of information we can incorporate into our evaluations. We can, for example, calculate the correlation statistics - both within the human set (to determine consistency within the set of human rankers) and between the set of human rankings and the machine generated statistic. The correlation statistic applicable to this data is Spearman's Rho [1], as the data is clearly non-parametric and on different scales - that is to say that any computer generated statistic is unlikely to map directly onto a 1-5 rating of interestingness. Nevertheless, if those videos rated highly by the computer are those videos rated highly by the human volunteers this is a positive result.

Spearman's Rho is a similar calculation to the product-moment correlation (sometimes called Pearson's), except Spearman's operates on ranked data. Given ranked data, Spearman's can be calculated using the following formula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Where n is the number of videos, and d is the difference between the matched pairs of ranks. Spearman's Rho can be tested for significance: for small values of n, r_s has a non standard distribution and specific tables must be used. For large ($n > 10$) values of n the following function of r_s follows approximately the distribution of a t-test statistic with $n-2$ degrees of freedom:

[1] These actors often look suspiciously like computer vision postgraduates.

$$t_s = \sqrt{\frac{n-2}{1-r_s^2}}$$

The resultant value t_s can be compared against any standard statistical tables for significance testing.

As well as the possibility of performing a range of statistical tests we have a wealth of qualitative information in the form of comments made by the subjects as they were ranking the dataset. These can help in instances where disagreement occurs - for example, in one outdoors scenario an object was reported as being highly interesting by several subjects, but the trajectory taken by that object was very dull. Inspection of their forms revealed that it was interesting because it was an ambulance.

3 The evaluative schema applied to the PETS dataset

A subset of the PETS2004 dataset was used in this study.[2] This consists of pedestrian footage filmed in a foyer situation, with actors performing various roles such as meeting, walking, fighting and browsing. Included in the dataset are various people we assume are bystanders. As we are only interested in evaluating high level classifications of behaviour (and not tracking), we only consider those videos for which ground truth has already been provided. We then exclude those agents whose trajectories are only partially covered by the video, and those agents who hover on the periphery. In short, we only analyse the main actors in each scene, and those bystanders whose trajectories are shown in full. This leaves us with a total of 23 agents from 12 movies. These are listed in detail, alongside an image showing the path of the trajectory, in Appendix A at the end of this paper.

The 12 original videos are used to produce 23 ($n = 23$) labelled videos. These were presented to 12 subjects ($n_s = 12$), who rated each on the 1-5 scale as detailed in Section 2. Spearman's Rho was calculated for each pair of human raters, giving a correlation matrix with 66 entries ($\frac{n_s^2 - n_s}{2}$). All of these correlations were positive, and 62 of the 66 were significantly positive at the 0.05 level. This means we can safely assume that the group of humans are in broad agreement about which clips are interesting.

It is interesting to take a closer look at the behaviour of those agents where the human rankers were in disagreement - where the standard deviation of the human scores is high. Some of these were due to partial trajectories, and to the inclusion of people such as ID1 from Walk3.mpg, who entered the scene then immediately turned around and left (we assume he was a passer-by, perhaps put off by the camera).

[2]This data comes from from the EC Funded CAVIAR project/IST 2001 37540

In particular, there are four cases in particular where the human rankings range from lowest (1) to highest (5) and it is worth investigating these in a little more detail:

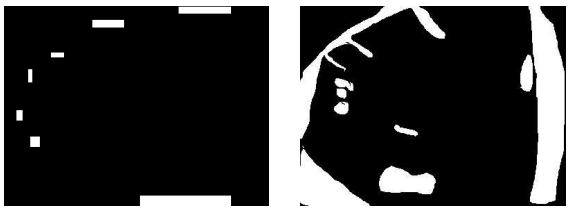
- **ID 0 from Walk1.mpg:** Standard Deviation = 1.07. In this movie clip, the agent walks out and waves at the camera, then leaves the scene by the same door they came in from. The actor in this clip is presumably signalling to the camera person that they are ready to go, although this was not clear from context.
- **ID 0 from Rest_SlumpOnFloor.mpg:** Standard Deviation = 1.48. In this movie, the agent walks out of the scene (the clip clearly starts before the actor is ready) then re-enters, crosses to the object on the left, then sits on the floor for a short while before leaving. Some of the subjects think that sitting on the floor was uninteresting.
- **ID 1 from Meet_WalkSplit.mpg:** Standard Deviation = 1.62. This clip and the following feature agents entering the scene from different doors, meeting in the middle, and then leaving from different doors. Comments by those subjects who rated these clips highly indicate that they thought a package was passed between the two actors - which would be suspicious given the instructions to subjects.
- **ID 3 from Meet_WalkSplit.mpg:** Standard Deviation = 1.56. See above.

That there was disgreement between the human subjects on some of the clips should not be seen as a drawback to this evaluative schema - indeed, one of the reasons for including a number of subjects is to allow for such differences and disagreements. These help provide a richer framework against which to evaluate our software.

4 A prototype *interesting event* detector

The interesting event detector we will use as a demonstration is inspired Dennett's large body of work on intentional explanation (for example, [2]), which describes different ways of thinking about and explaining the behaviour of agents. It is hoped that this system, when completed, will provide a new way of thinking about the problem of behaviour modelling in the surveillance domain. However, it is still at the prototype stage and we will just sketch an overview of its operation here.

Our system tries to work out where the agent might be heading, and combines this hypothesis with a simple model of the way in which people intentionally navigate towards the geographical goals in a scene. In our original formulation, the hypothesis about where the agent may be heading



(a) Exits (b) Obstacles

(c) Scene

Figure 1. The exit model, obstacle model and scene.

is built up using information from a person tracker [5], an obstacle model and an exit model, but in the current implementation we simply use the ground truth information provided (specifically, the position of the object centroid) for the agent's position ($\mathbf{x}$). We apply a Kalman filter to this and store the directional component to obtain an estimate of direction of travel θ.

All calculations are carried out in the image plane, and we make the simpifying assumption that the wide-angle lense used to capture the PETS2004 datasets will not have a significant effect on our calculations.

Central to our approach is the concept of a *goal*. We define goals as places where the agent can leave the scene - doors and exits - or to borrow terminology from Ellis and Xu [11] "Long-term Occlusions" or "Border Occlusions". In the types of pedestrian scene typically subject to surveillance, these are the goals of the agents therein - in a car park, the pedestrian goals are either their cars or the door; in a general pedestrian scene like a shopping mall or the foyer of a research institute, the goals are the exits of the scene, and perhaps some other form of attraction such as an information desk or an ATM machine. The goals for any particular scene can be learned, if enough example footage is present. In the current experiment the PETS2004 dataset (which features a number of short videos) does not provide the body of data required for learning to take place, and so the exit model was hand crafted. This consists simply of rectangular boxes representing each exit. The obstacle model is similarly hand crafted, but for computational reasons does not have to be regular and is simply stored as a bitmap. Figure 1 shows these models and an image of the scene.

Our central assumption is that people move consistently towards their goal. If there is an obstacle between the agent and their goal, virtual "sub-goals" are constructed in places where the agent might be able to see more of the scene than they currently can - thus, sub-goals are constructed on the edge of obstacles, in places where the agent would be able to see further around the obstacle. From each sub-goal, we compute which goals would be visible if the agent were at that point, and also any sub-sub-goals. And from each sub-sub-goal, we compute which further goals would be visible. Thus for each goal $\mathbf{x_g}$ within the scene we can determine whether that goal is directly visible, or whether it would be visible by turning a corner, or whether it would be visible by turning two corners (in the current implementation we stop computation at two levels of sub-goal analysis). This takes the form of a label - $Label(\mathbf{x_g})$ - which can have the values V, for those goals which are directly visible; N, for those goals which are not visible at all, and $S1$ or $S2$ for those goals which are accessible via a sub-goal or two.

Indeed, there are four possible relationships between an agent and each goal for each frame, which can be determined from the label of the pixel at the position of the goal $Label(\mathbf{x_g})$, and the angle ϕ, which is the angle subtended by a line between the position of the goal $\mathbf{x_g}$, the position of the agent $\mathbf{x}$, and the agent's current direction estimate θ. These are:

1. A: The goal is directly visible: $Label(\mathbf{x_g}) = V$; and the agent is heading towards it $1 > \phi > -1$. $g2$ is in this state in Figure 2.

2. D: The goal is directly visible to the agent: $Label(\mathbf{x_g}) = V$; but they are heading away from it: $\phi > 1$ *or* $\phi < -1$. $g4$ is in this state in Figure 2.

3. N: The goal is not visible to the agent: $Label(\mathbf{x_g}) = N$ (it is on the other side of an obstacle, and is not reachable by means of a sub-goal) . $g3$ is in this state in Figure 2.

4. Sn: The goal is visible to the agent, but only via a sub-goal ($S1$) or a sub-sub-goal ($S2$): $Label(\mathbf{x_g}) = Sn$. $g1$ is in state $S1$ in Figure 2.

Given these frame-by-frame classifications for each goal, we can build an idea of how likely - or rather, unlikely - that goal is as an explanation for the trajectory as a whole. This is done by associating a cost with certain state transitions. The diagram in Figure 3 shows costs associated with transitions in the current model.

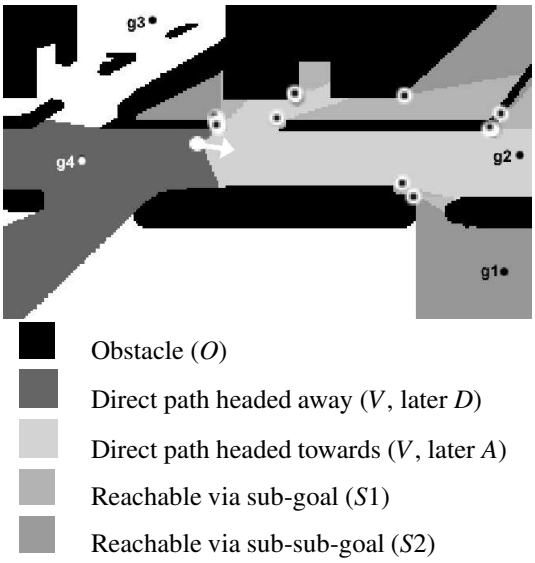


Figure 2. An example of the sub-goal algorithm in action in an outdoor pedestrian scene. The agent is represented by a white dot and a white arrow (corresponding to its velocity vector); white dots with black centres are sub-goals; the obstacle model is shown in black; areas which are not visible (either directly or via a sub-goal or two) in white; areas shaded very light grey represent areas directly visible and headed towards; darker shades of grey represent areas of the scene accessible only via sub-goals or sub-sub-goals; very dark grey represents areas directly visible but not within the angle of vision; $g1$, $g2$, $g3$ and $g4$ are example goals referred to in section 4.

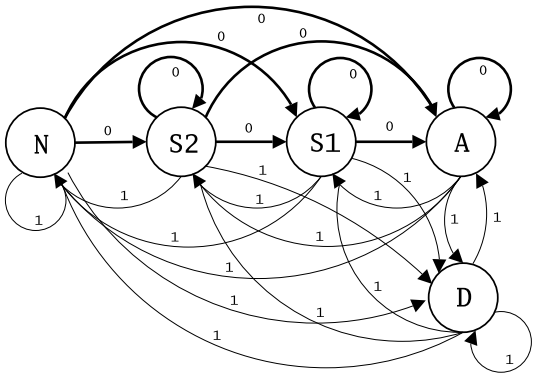


Figure 3. State transition diagram indicating the cost of each transition. Those transitions which are *free* (drawn with thick lines) are those associated with progress towards the particular goal; those with a cost are those associated with movement away from the goal

Applying costs as laid out in Figure 3 provides us with a cost for each goal within the scene, and that cost can be thought of as representing the number of frames in which the agent's behaviour is inconsistent with travel towards that particular goal. These costs are then divided by the total length of the trajectory to provide a statistic which is comparable across agents. Finally, we need to simplify matters and provide a single cost for each actor. If the system were fully recursive and we were able to work out the final exit for each agent, the cost associated with the final exit would be an alternative measure. However, in the current dataset there are some trajectories which finish whilst the agent is still in view of the camera, making this statistic unreliable. The highest cost or average cost would be inappropriate, as it is possible for perfectly uninteresting trajectories to avoid one or more exits completely; these would have very high costs indeed - this would also affect any attempt to use the average cost or some other aggregate score over all goals. Therefore, in the current situation, the best choice for a single cost is the lowest cost.

We can think of the Cost% statistic as representing the percentage of frames in which the agents' travel was inconsistent with motion towards *their most likely* goal. Cost% scores for the PETS2004 dataset are set out in full in Appendix A. The next section of this paper discusses ways in which our Cost% statistic can be compared with the "ground truth" scores discussed in section 3. It is worth noting that with simple scenes without obstacles, the algorithm just described simplifies to straightest path and the sub-goal mechanism does not make any difference to the output. For several of the simpler trajectories in the PETS2004 dataset this was indeed the case, and the power of the approach would be better demonstrated in a more complicated scene with multiple obstacles. That said, the results on the PETS2004 dataset are still promising and worth discussion.

5 The prototype evaluated

The question we now have to address is how well our prototype results agree with the results of the subjects detailed in Section 3. Firstly, we can calculate Spearman's Rho - r_s - the correlation coefficient, between the computer generated Cost% statistic and the human subjects, and between the Cost% and the human mean and median (making the assumption that it is appropriate to reify the averages in this way). The correlation statistic r_s and the t-statistic t_s are

Correlation with	r_s	t_s
H1	**0.639**	3.807
H2	**0.679**	4.234
H3	**0.408**	2.05
H4	*0.353*	1.729
H5	**0.507**	2.692
H6	**0.453**	2.329
H7	0.277	1.319
H8	0.292	1.4
H9	*0.386*	1.917
H10	0.319	1.542
H11	**0.47**	2.439
H12	**0.626**	3.676
Median Human	**0.607**	3.499
Mean Human	**0.639**	3.810

Table 1. Correlation statistics for the Cost% score against each individual subject and the human averages. Those values which are statistically significant at the 0.05 level are highlighted in boldface, and those which are significant at the 0.1 level but not the 0.05 in italics.

set out in Table 1.

The significance levels for t_s with $n = 23$ are 1.721 at the 10% level and 2.080 at the 5% level. As is clear from Table 1 the correlation with the average human is statistically significant. In Figure 4 we have drawn the graph of Cost% and the median human score by video clip (sorting the video clips by median) it is clear that those clips rated highly by humans generally scored highly on the machine generated statistic as well. This graph also enables us to see the anomalous cases clearly.

The spike labelled A in Figure 4 corresponds to Meet_WalkTogether1.mpg id 2. In this clip, the agent enters from one side, meets someone, changes direction and heads towards an exit - he does not actually exit the scene, but turns around and comes a short way back into the foyer before the video cuts off. We assume that this is an artifact of video editing - it is definitely strange behaviour if not. Given that our system can be thought of as providing a measure of behaviour consistency, it is acceptable for it to pick up on such artifacts.

The spikes labelled B and C in Figure 4 correspond to Meet_WalkSplit.mpg. These trajectories are both quite complicated, involving first moving towards the other agent in the scene and then moving towards an exit (different exits in each case). This is an aspect of our software that we hope to address in future - specifically, making other tracked agents within the scene legitimate goals in themselves. This

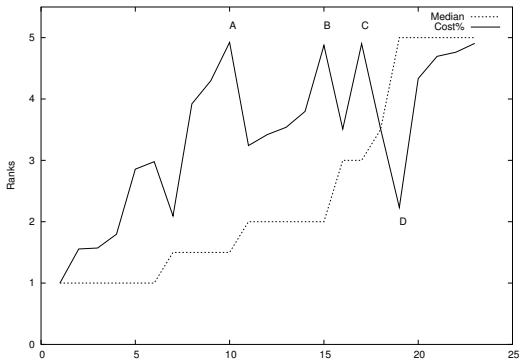


Figure 4. A comparison of the Cost% statistic and the median human rankings. Cost% has been scaled, in order to place both outputs in the same range (1-5). The x-axis values are ordered by median.

is one of the clips highlighted for attention in Section 3 as being a video with high variance amongst human rankers.

The trough labelled D in Figure 4 corresponds to id 6 in Fight_RunAway1.mpg. This agent enters the scene moving quite rapidly, has a play-fight with another agent (lasting just a few seconds), then runs across to the exit opposite. His trajectory is essentially a straight line with a slight kink in the middle, and as our software operates solely on individual trajectories does not pick up on this behaviour. An extension to our system which might cope with this would be to take into account the relative positions of other agents in the scene.

6 Conclusions

Evaluation in Computer Vision should be about more than merely x,y,t. And even when evaluating something as simple as x,y,t, it has been suggested [8] that reliance on just one estimate is unwise. As we produce more complicated systems, performing higher level cognitive tasks than "simple" classification or location, we need more complicated, higher level evaluative techniques. If a system is presented as a general-purpose surveillance system, or an "Interesting Event Detector", then it should be evaluated as such. Specifically, it should be evaluated in such a way that the opinions and prejudices of the designers cannot affect the evaluation. Evaluation by accident - simply noting that the events detected seem to be odd - is not good enough. Evaluation by actor - by engineering test cases which involve people behaving strangely - is suspect, and evaluation based upon the opinion of the system author is also unsatisfactory. In this paper, we have presented a novel approach

which uses a group of naïve subjects who together provide a rich background against which the performance of an algorithm can both be measured statistically and compared qualitatively.

The software outlined in this paper is also novel, in that it adopts a high level intentional analysis of what is essentially quite simple behaviour. Previous work consists of analyses of the resultant behaviour: the fact that people follow similar trajectories across a scene [4] is because they have similar goals; the fact that paths can be approximated by trajectory analysis [6] is because paths join two goals. This work attempts instead to analyse the cause of the behaviour – the goals – directly. The initial results presented here are promising, and show that in principle such an analysis could be used in a practical situation to provide a filter on surveillance data.

References

[1] Clarke G.M. and Cooke D. *A basic course in statistics: third edition*. Edward Arnold, London, 1992, 3 edition.

[2] Dennett D.C. 'True believers: The intentional strategy and why it works.' In: W.G. Lycan (editor), *Mind and Cognition: A Reader*, pp. 150–167. Blackwell, Cambridge, MA, 1990.

[3] Jan T., Piccardi M. and Hintz T. 'Detection of suspicious pedestrian behavior using modified probabilistic neural network.' In: *Proc. of Image and Vision Computing*, pp. 237–241. Auckland, New Zealand, 2002.

[4] Johnson N. and Hogg D.C. 'Learning the distribution of object tractories for event recognition.' *Image and Vision Computing*, Vol 14(8), pp. 609–615, 1996.

[5] Magee D.R. 'Tracking multiple vehicles using foreground, background and shape models.' *Image and Vision Computing*, Vol 22, pp. 143–155, 2004.

[6] Makris D. and Ellis T. 'Spatial and probablistic modelling of pedestrian behaviour.' In: *Proc. British Machine Vision Conference*, pp. 557–566. Cardiff, UK, 2002.

[7] Morris R.J. and Hogg D.C. 'Statistical models of object interaction.' *International Journal of Computer Vision*, Vol 37(2), pp. 209–215, 2000.

[8] Needham C.J. and Boyle R.D. 'Performance evaluation metrics and statistics for postitional tracker evaluation.' In: *Proc. International Conference on Computer Vision Systems*. Austria, 2003.

[9] Stauffer C. and Grimson E. 'Learning patterns of activity using real-time tracking.' *Pattern Analysis and Machine Intelligence*, Vol 22(8), pp. 747–757, 2000.

[10] Troscianko T., Holmes A., Stillman J., Mirmehdi M., Wright D. and Wilson A. 'What happens next? the predictability of natural behaviour viewed through CCTV cameras.' *Perception*, Vol 33(1), pp. 87–101, 2004.

[11] Xu M. and Ellis T. 'Partial observation vs. blind tracking through occlusion.' In: *Proc. British Machine Vision Conference*, pp. 777–786. Cardiff, UK, 2002.

A Table of agents and results

Filename	Id	Image	Description	Cost% (Scaled)	Human mean	Human SD	Human Median
Walk1.mpg	0		Walks in, waves at camera, goes back through same door	37.8 (3.52)	3.33	1.07	3.5
Walk1.mpg	1		Walks slowly across scene	8.57 (1.57)	1.25	0.45	1
Walk3.mpg	1		Walks out, turns around, walks back through same door	38.1 (3.54)	2.08	1.16	2
Walk3.mpg	2		Walks slowly across scene	16.4 (2.09)	1.58	0.67	1.5
Meet_WalkTogether1.mpg	1		Enters, meets, shakes hands, changes direction, exits	49.46 (4.3)	1.92	1.16	1.5
Meet_WalkTogether1.mpg	2		Enters, meets, shakes hands, changes direction, exits	43.82 (3.92)	1.92	1.16	1.5
Rest_FallOnFloor.mpg	2		Enters in a wobbly fashion, falls over, gets up and leaves	58.6 (4.91)	4.67	0.65	5
Rest_SlumpOnFloor.mpg	0		Leaves scene, re-enters, slumps on floor, leaves scene again	58.52 (4.9)	3	1.48	3
Meet_WalkSplit.mpg	1		Walks towards person, shakes hands, turns, leaves scene	58.13 (4.88)	2.5	1.62	2
Meet_WalkSplit.mpg	3		Walks towards person, shakes hands, turns, leaves scene	36.31 (3.42)	2.33	1.56	2
Meet_Crowd.mpg	0		Walks in straight line across scene	8.33 (1.56)	1.33	0.78	1
Meet_Crowd.mpg	1		Walks in straight line across scene	11.95 (1.8)	1.5	0.67	1
Meet_Crowd.mpg	2		Walks in relatively straight line across scene	27.86 (2.86)	1.5	1	1
Meet_Crowd.mpg	3		Walks in relatively straight line across scene	29.67 (2.98)	1.58	1.16	1
Fight_RunAway1.mpg	6		Walks in, fights, runs out	18.5 (2.23)	4.75	0.62	5
Fight_RunAway1.mpg	7		Hangs around, Walks in, fights, runs out	56.44 (4.76)	4.67	0.65	5
Fight_OneManDown.mpg	4		Walks in, fights, runs in circles, runs out	50 (4.33)	4.75	0.62	5
Fight_OneManDown.mpg	5		Enters, gets fought with and knocked over, leaves	55.43 (4.7)	4.33	1.15	5
Browse_WhileWaiting2.mpg	0		Wanders aimlessly	41.97 (3.8)	2.08	0.9	2
Browse4.mpg	1		Wanders aimlessly	33.62 (3.24)	1.92	0.9	2
Browse4.mpg	2		Walks directly across scene	0 (1)	1.17	0.39	1
Browse2.mpg	1		Walks in, waves at camera, leaves	37.66 (3.51)	2.75	0.97	3
Browse2.mpg	3		Wanders towards bookshelves, browses, leaves	58.87 (4.92)	1.67	0.78	1.5